One-class classification, Multiclass classification and Multilabel classification

Sunaina, Bibhu Prasad Behera

National Institute of Science Education and Research (NISER),
Bhubaneswar

April 21,2023

One-class classification and multiclass classification

Part I

CLASS CLASSIFICATION AND MULTICLASS CLASSIFICATION

Introduction

The problem of classification can be succinctly explained as trying to assign a new object into a set of classes. These classes can be either mutually exclusive or not. The number of these classes also varies from either one to more than one. For such segregation, a set of rules exist for each class but the elucidation of such rules is highly dependent on the limited data for each class. In a real-world dataset, this task is further complicated by the existence of missing data for each class and the resulting imbalance that is observed.

Similarly, for classifying a real world dataset, it is assumed that the new object, if it's closer to an object of some class, it actually belongs to that class. Although this might seem obvious, it is an important assumption and may have many test cases where it might give an unexpected result.

- When mutually exclusive classes are present, the classification of one class is called One-class classification (OCC) and for more than one class, it is called multi-class classification.
- When the classes are not mutually exclusive, it becomes a multi-label problem that has been described in detail in further sections.

INTRODUCTION

A REAL WORLD EXAMPLE: MALICIOUS BEHAVIOUR DETECTION IN NETWORKS

Through analysing the network traffic data, we can classify malicious behaviour as opposed to normal behaviour of the data. This can be a one class problem for a network which has a low number of malicious attacks as seen in Fig. 1. When the frequency of malicious attacks is as high as normal behaviour, a multi-class approach can be used.

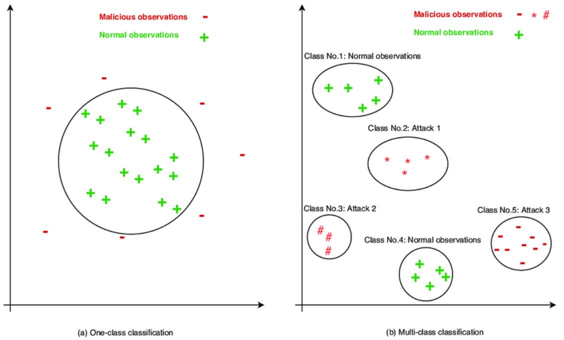

Figure. 1: Comparison between one-class classification and multi-class classification

Approaches towards solving the one-class classification problem

One-class Support Vector Machine

One-class support vector machines (OC-SVMs) are trained on a dataset containing only one class. The new objects are either classified as belonging to that class or as an outlier. Majorly, there have been two different algorithms for it:

- Similar to the linear SVM for multiple classes, the data points are projected to a higher dimensional feature space. As opposed to using a hyperplane, Tax et al's paper uses a "hypersphere" to separate outliers. This hypersphere is minimized to contain those datapoints that belong to the class whereas the outliers are kept outside the sphere. There are additional parameters that help in adjusting how many data points are allowed in the sphere.
- Schölkopf et al's paper on the other hand applies a hyperplane similar to linear SVM to separate most of the datapoints away from the origin with maximum distance between them as seen in Fig. 2.

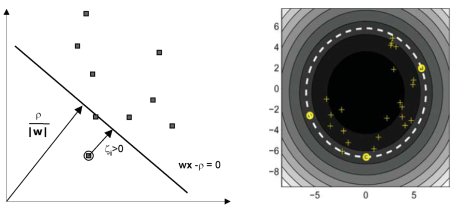

Figure. 2: Left: Hyperplane Right: Hypersphere

Approaches towards solving the one-class classification problem

A needle in a haystack

The loss function depending on the need is varied. For instance, you want to develop a search engine. Out of millions of web pages, on the basis of a few words, you only want to output the relevant papers. As an analogy, you want to find a needle in a haystack. Especially for a search engine, you don't want ***any*** irrelevant pages on the first search result page. In this case, the loss function is increased inside the hypersphere from the center and kept constant outside it where the outliers are present. This accomplishes two things –

- The hypersphere is kept as low as possible.
- It doesn't take the outliers into account at all.

Whereas if you're fine with outliers in the classification, the loss function is kept constant inside the hypersphere and it increases outside it.

Approaches towards solving the one-class classification problem

Tackling imbalanced datasets

One-class classifiers are especially useful in tackling imbalanced datasets where there is a majority class and a minority class. Although they are defined for a single class, there have been studies where they have been applied on imbalanced datasets. This occurs in three ways –

- Application of one class classifier on the majority class
- Application of one class classifier on the minority class
- Application of one class classifier on both the classes

This will help to detect outliers in the data and to better understand the underlying description for the majority class.

Multiclass classification

Problems in solving a multiclass problem

One cannot simply apply many of the binary classifers to a multi-class problem. Specifically, logistic regression and support vector machines are designed for binary classification problems and need to be iteratively applied for the different classes to make use of them in a multi-class problem. There are three simple ways to do this:

- One-vs-All
- All-vs-All or One-vs-One

Multiclass classification

One-vs-All

Assuming we need to classify an object into N classes, this approach generates N different binary classifiers – one for the classification of each class. For example, the classes can be "apple", "pear" and "strawberry". So there would be three binary classifiers:

- Apple vs the rest: Positive instances of apple are taken as correct classification whereas negative instances includes the other two.
- Pear vs the rest: Positive instances of pear are taken as correct classification whereas negative instances includes the other two.
- Strawberry vs the rest: Positive instances of strawberry are taken as correct classification whereas negative instances includes the other two.

For each class, a probability score of an object belonging to that class is produced. The argmax of these scores are taken and the model with the highest argmax is taken. This is the default choice of the algorithsm provided in the sci-kit learn library.

Multiclass classification

All-vs-All

Assuming we need to classify an object into N classes, this approach generates $\frac{N \times (N-1)}{2}$ different binary classifiers – one for the classification of each class with respect to every other class. Using the same example, there would be three binary classifiers:

- Apple vs Pear: Positive instances of apple are taken as correct classification whereas negative instances includes objects classified as pear.
- Pear vs Strawberry: Positive instances of pear are taken as correct classification whereas negative instances objects classified as strawberry.
- Strawberry vs Apple: Positive instances of apple are taken as correct classification whereas negative instances include objects classified as apple.

For each object, probabilities for each class are summed up and the class with the largest probability score is used for classification.

Multiclass classification

Comparison between One-vs-all and All-vs-All

All-vs-All approach creates a higher number of models for the different classes as opposed to one-vs-all in Fig. 3 3. But at the same time, one-vs-all calculates independent probabilities of an object belonging to a class. The final probabilities will not sum up to 1 and they may need to be calibrated again using other methods.

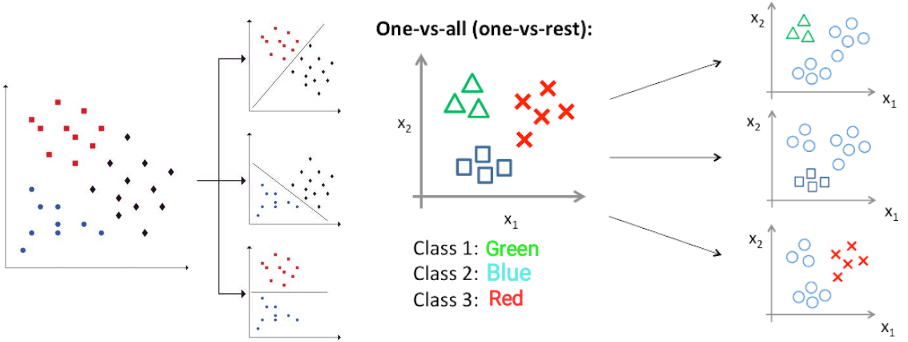

Figure. 3: Left: All vs all classification, Right: One vs all classification

Conclusion

- One-class classification problems are a lot harder to solve than multi-class classification problems. Or in other words, it's harder to describe an object and derive its underlying rules as compared to differentiating it against its negative instance.
- The approaches to dealing with both the problems are different. Most binary classifiers have been amended to deal with multi-class problems.
- Both classifications have different test uses. One-class classification can be used for outlier and novelty detection, in species distribution models where there is only one class or in imbalanced datasets.
- Multi-class classifications can be used in almost any case where the datasets are balanced.

References I

- Multiclass Classification - MIT Class 9.520
- COMP 642 - Machine Learning
- Rabbani, M., Wang, Y., Khoshkangini, R., Jelodar, H., Zhao, R., Bagheri Baba Ahmadi, S., Ayobi, S. (2021). A Review on Machine Learning Approaches for Network Malicious Behavior Detection in Emerging Technologies. Entropy, 23(5), 529. https://doi.org/10.3390/e23050529
- Crammer, Koby Chechik, Gal. (2004). A Needle in a Haystack: Local One-Class Optimization. Proceedings of the Twenty-First International Conference on Machine Learning (ICML): 2004. https://doi.org/10.1145/1015330.1015399
- Tax, D.M., Duin, R.P. Support Vector Data Description. Machine Learning 54, 45–66 (2004). https://doi.org/10.1023/B:MACH.0000008084.60811.49
- Schölkopf, B. and Williamson, RC. and Smola, AJ. and Shawe-Taylor, J. and Platt, JC. Support Vector Method for Novelty Detection. Advances in Neural Information Processing Systems, 582-588 (2000). 10.1.1.675.575

Part II

Multilabel classification

Multilabel Classification

Definition of Multilabel classification

Multilabel classification is a type of machine learning problem where an input can belong to more than one class or category. In other words, it is a classification task where each data point can be assigned to one or more label at the same time.

More about Multilabel Classification

- Multilabel classification assigns one or more labels to each data point.
- Commonly used in image classification to identify multiple objects within an image.
- One-vs-rest and multioutput classifiers are two common approaches to multilabel classification.
- Evaluation metrics include precision, recall, and F1-score, calculated separately for each label.

This machine learning technique has multiple applications in a lot of different industries. it can be used in various works starting from image tagging to Automatically detect adverse drug reaction from the text provided.

By assigning various tags and labels, we can do the following with multilabel classification:

- Creating user profiles
- Building a recommendation system
- Social media targeting

Multilabel Classification

Where Multilabel Classification is useful

- In **Text classification**, multilabel classification can be used to identify the topics, themes, or sentiments expressed in a piece of text.
- In **Image tagging**, multilabel classification can be used to identify the objects, scenes, and attributes present in an image, such as "cat," "beach," and "sunset."
- In bioinformatics, multilabel classification can be used to predict the **functions and interactions of proteins** based on their amino acid sequences or structures.
- Multilabel classification is also used in **Recommendation systems**, where it can be used to predict the user's preferences for multiple items or products simultaneously.
- Another application of multilabel classification is in **Social media analysis**, where it can be used to identify the topics and sentiments expressed in a collection of tweets or posts.

Multilabel Classification

Where Multilabel Classification cannot be used?

Multilabel classification is not suitable for all types of data. In cases where the labels are mutually exclusive and cannot occur together. This is because multilabel classification is designed to handle situations where each data point can belong to multiple labels simultaneously. So if a data point belongs to 1 or 2 labels then this classification technique will not work.

If the labels are mutually exclusive, meaning that a data point can belong to only one label at a time, then a different classification approach such as binary or multi-class classification should be used instead.

Additionally, if the label space is small or simple, and each data point has only one or a few labels, then multilabel classification may not be necessary or may not provide significant advantages over simpler classification techniques.

Multilabel Classification

Challenges of Multilabel Classification

High Dimensionality of the Label Space:

- The label space can be very large and complex, which can make it difficult to train a model that can accurately predict all possible combinations of labels.
- It is important to carefully consider the label space and identify any redundant or irrelevant labels.

Evaluation of the Model Performance:

- Evaluating the performance of a multilabel classification model is more complex than in other classification tasks because there are multiple labels involved.
- Common evaluation metrics for multilabel classification include precision, recall, and F1 score.
- Other metrics such as Hamming loss, Jaccard similarity, and micro/macro averaged precision and recall can also be used.

Class Imbalance Problem:

- Some labels may occur more frequently than others, which can create a class imbalance problem.
- The model may be biased towards the more frequent labels and may struggle to predict the less frequent ones.
- Techniques such as oversampling, undersampling, or the use of class weights can be used to balance the distribution of labels in the training data.
- Class imbalance problem in multilabel classification can lead to biased models and a high number of false negatives, which can be problematic in applications where missing a label can have serious consequences.

Differences between multiclass classification and multilabel classification

- In multiclass classification, each instance is assigned to only one class, while in multilabel classification, each instance can be assigned to one or more labels simultaneously.
- Multiclass classification is used when the target variable has multiple classes, but only one class is applicable to each instance, while multilabel classification is used when the target variable has multiple labels, and more than one label can be applicable to each instance.
- In multiclass classification, the classes are mutually exclusive, while in multilabel classification, the labels may or may not be mutually exclusive.
- Multiclass classification is simpler to evaluate as it only requires metrics such as accuracy or confusion matrix, while multilabel classification requires more complex evaluation metrics, such as precision, recall, and F1-score for each label.

Conclusion

In summary, multilabel classification is a type of machine learning problem that is used in various fields to assign multiple labels or categories to an input. It is not suitable for all problems, and its complexity requires careful consideration of the appropriate techniques for data preparation, modeling, and evaluation.

References I

- **TASQ** https://www.tasq.ai/blog/multi-label-classification/
- **Scikit-learn** /multiclass.htmlhttps://scikit-learn.org/stable/modules/multiclass.html
- **Machine learning mastery** https://machinelearningmastery.com/multi-label-classification-with-deep-learning/
- **Geeksforgeeks** https://www.geeksforgeeks.org/an-introduction-to-multilabel-classification/https://www.geeksforgeeks.org/an-introduction-to-multilabel-classification/
- **Scikit** https://scikit-learn.org/stable/modules